DOCUMENT RESUME

ED 329 587 TM 016 231

AUTHOR Roid, Gale H.
TITLE Historical Continuity in Intelligence Assessment: Goals of the WISC-III Standardization.
PUB DATE Aug 90
NOTE 13p.; Paper presented at the Annual Meeting of the American Psychological Association (98th, Boston, MA, August 10-14, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Children; Cognitive Measurement; Diagnostic Tests; Educational History; Elementary Secondary Education; *Intelligence Tests; Item Bias; Psychological Testing; Social Science Research; *Standardized Tests; *Test Construction; Trend Analysis
IDENTIFIERS Revision Processes; *Standardization; *Wechsler Intelligence Scale for Children (Revised)

ABSTRACT

The development and standardization of the third edition of the Wechsler Intelligence Scale for Children-Revised (WISC-R) are described. The existence of documented historical trends was a significant factor in the project to revise the WISC-R, which is the most widely used intelligence battery in North America. Development of the third edition (WISC-III) began in the mid-1980's. An extensive array of research studies was conducted to enhance the subtest composition and characteristics of the third Wechsler factor. Improvements to the subtest included: (1) expanding the object assembly subtest; (2) adding modernized artwork; (3) improving the coding subtest; (4) extending the normative range of subtests; (5) minimizing item bias; and (6) beginning development of supplementary materials and tests. The major features of its predecessor were kept in the WISC-III for historical continuity. In addition, greater downward age compatibility was created because of the recent revision of the preschool scales. One flowchart illustrates the Wechsler Model of Intelligence. A 30-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED329587

HISTORICAL CONTINUITY IN INTELLIGENCE ASSESSMENT: GOALS OF THE WISC-III STANDARDIZATION

by

Gale H. Roid
Oregon Department of Education
and
Assessment Research
Salem, Oregon

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
GALE H. ROID
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

August, 1990

Paper presented as part of the symposium "Development of the Wechsler Intelligence Scale for Children -- Third Edition" at the annual meeting of the American Psychological Association, Boston, August, 1990. Inquiries concerning this paper should be directed to Dr. Roid at the Department of Education, 700 Pringle Parkway S.E., Salem, OR 97310.

TM016231

HISTORICAL CONTINUITY IN INTELLIGENCE ASSESSMENT: GOALS OF THE WISC-III STANDARDIZATION

Introduction

Historical accounts of the development of measures of intelligence and cognitive ability have been provided by Carroll (1989), Jensen (1980), Sattler (1988) and others. A comprehensive review of international data on changes in average intelligence quotients (IQ) within countries was provided by Flynn (1987). Flynn, for example, found significant gains in nonverbal cognitive scores (e.g., Raven's matrices, Wechsler scores) of 5-25 points in a single generation within countries having high quality data. Whether these gains represent true enhancements in multifaceted intelligence or whether they represent gains in a more specific "abstract, decontextualized problem-solving ability" (Flynn, 1987), they are remarkable and consistent findings. Regardless of controversies over the true nature of intelligence as unitary, hierarchical or multifaceted, the existence of standardized measures comparable over time provides valuable data for studying human cognitive functioning.

The existence of documented historical trends was a significant factor in the project to revise the Wechsler Intelligence Scale for Children -- Revised (WISC-R). The WISC-R (Wechsler, 1974) is the most widely used intelligence battery in North America, and was a revision of the original scale (WISC) published in 1949. The development of a third edition of the battery began in the mid-1980's, and intensified in the spring of 1987 when the actual time-table began for the publication of the Third Edition (WISC-III). As described below, the consensus of many experts in educational and psychological measurement was "Don't change it!" Providing continuity and the possibility of studying historical trends were major factors in the design of WISC-III.

The following were major elements of the planning for WISC-III, and each will be discussed in turn:

- Continuity and historical trends in intelligence theory.
- Enhancing the factor structure of WISC-III.
- Improvements and modernization of test materials, subtests and test-administration features.
- Minimization of potential item bias.
- Development of materials to supplement the battery -- a co-normed achievement test and computer-interpretive software.

Continuity and Theories of Intelligence

A number of new theories of intelligence have emerged during the last two decades. The prominent work of Sternberg (1984), Gardner (1982), Das, Kirby and Jarman (1975), Cattell and Horn (1981), Gustafsson (1984) and others have opened the door to a major re-examination of our understanding of intelligence and mental ability. The growth of cognitive psychology has stimulated numerous attempts to analyze the cognitive components of the tasks in intelligence batteries (e.g., Campione & Brown, 1978; Carroll, 1987; Embretson, 1985; Hunt, 1987; Jensen, 1987; Pellegrino & Glaser, 1979). Rather than injecting extensive new theories into the Wechsler scales, the decision was made to let other theories emerge in alternative published batteries. Thus, the comprehensive battery by Elliott (1990), the Differential Ability Scales was published and a new instrument by Sternberg (in press) developed. This allowed the third edition to maintain continuity with earlier Wechsler editions.

A more compelling reason to maintain the theoretical underpinnings of the Wechsler scales was the enduring importance of g theory. Despite the proliferation of theories of intelligence, a number of common elements have survived since the pioneering work of Spearman (1927), Vernon (1950) and others. The concept of a general factor, g, remains strong in terms of theory (e.g., Gustafsson, 1984) and empirical evidence has continued to accumulate on its importance to practical criteria such as job performance (Jensen, 1987) and scholastic achievement. An example of current research that assumes the existence of g and numerous cognitive components is the work of Carroll (1989). Carroll is currently exploring 461 correlation matrices from a multitude of cognitive batteries using hierarchical factor analysis. In Carroll's method, each subtest from each of the major cognitive batteries is assumed to have a loading on both a primary factor (e.g., verbal comprehension) and a general factor. Each published test battery has its own mix of primary and specific factors, but Carroll's work may show the degree of overlap in the general factors among the prominent test batteries.

The Wechsler Model

The factor model underlying the WISC-R, and planned for WISC-III, is shown in Figure 1. As established by numerous factor analyses (e.g., Kaufman, 1979), WISC-R has typically shown three factors, Verbal Comprehension, Perceptual Organization and a third factor labeled "Freedom from Distractibility" by Kaufman.

Numerous studies of the third factor have appeared in the last decade, and the exact nature of this factor remains elusive. We have chosen to refer to the factor as a combination of attention and short-term memory, recognizing that any label is tentative at this time (see discussion of new subtest Symbol Search below).

Insert Figure 1 about here

Enhancing the Third Factor in WISC-III

In preparation for the revision of WISC-R, in-depth professional reviews were solicited from several leading experts in clinical and school psychology. These reviewers requested that the two major factors of the WISC-R be continued with some enhancement to the third factor (which was not to be elevated to the status of an IQ score like the Verbal IQ and Performance IQ). Based on these recommendations, the results of a telephone survey of WISC-R customers, and extensive review of the factor-analysis literature, the search for one or two additional subtests was begun during pilot testing (1987-88).

Reviews of the literature on the WISC-R third factor pointed to the possibilities of enhancing the memory, attention-span and/or numerical-ability facets of this complex factor. The research on memory-scanning abilities by Sternberg (1966) suggested an experimental subtests called "Symbol Search." The task include a long list of target sets of two symbols (non-alphabetic characters such as a circle with a superimposed "X") that must be identified as present/absent in a scanning-set of five symbols. The task is timed and the score is the number of correct identifications within the time limit. A parallel subtest using conventional alphabetic letters was also created for pilot testing.

A second experimental subtest was suggested by the work of Spring and Farmer (1975) -- Rapid Digit Naming. This research suggested that poor readers and dyslexics (who typically do poorly on Factor 3) are disrupted in their ability to repeat a sequence of single-digit numbers in rapid succession. A parallel subtest was also constructed using pictures of animals and objects familiar to children.

Pilot Testing and Tryout Data

Pilot testing on 119 children (approximately 40 in each of three age groups) was conducted with experimental subtests and original subtests (expanded with new items). In addition to important preliminary item analyses, these data were subjected to various factor analyses. The new symbol search subtest proved to have the most consistent pattern of correlations and factor solutions with the Coding, Digit Span, and Arithmetic subtests.

A extensive pilot experiment was later conducted on various parallel forms of Symbol Search to establish the best size and type of target and scanning sets (variations from one to two targets with three to five geometric-shapes or non-letter symbols in the scanning set). A final version of the new subtest was then included in the major tryout of WISC-III in 1988. The tryout

sample of 300+ children was administered original subtests (with new items) including enhancements to the Coding (more items and larger, more spacious printing) and Arithmetic subtests (experimentation with the reduction of memory load by presenting all Arithmetic items on cards for the child to read). Also included in the tryout was a new, experimental Visual Memory subtest. The new memory subtest was intended to supplement the memory facet of Factor 3. Results continued to point to the Symbol Search subtest, however, as the new subtest of choice.

In summary, an extensive array of research studies were conducted to enhance the subtest composition and characteristics of the third Wechsler factor.

Improvements to Subtests

Several sources of information were integrated into plans for the improvement of WISC-R subtests. Extensive files of user inquiries, a telephone survey of customers (using a structured questionnaire), suggestions from numerous consultants and experts, detailed lists of improvements from psychologists who do high-volume testing, and reviews of the literature on WISC-R provided a wealth of information to guide subtest improvements.

Object Assembly

It is well-known that the Object Assembly subtest of WISC-R is the Performance subtest with the lowest reliability at certain age levels. Also, the printing and production of accurate puzzle pieces is an expensive and demanding process. For these reasons an extensive search was made of the literature on spatial ability, neuropsychological assessment, and previous paper-and-pencil measures of perceptual-organization abilities (e.g., Thurstone, 1938). An experimental subtest, labeled "Picture Assembly" was drafted and given as part of the Pilot Test in 1987-88. The items were line drawings of familiar objects which were "cut apart" and printed in spaced and rotated arrangements which the subject was to visually reassemble. Unfortunately, the experimental subtest did not produced a sufficient degree of enhanced reliability over that achieved by an expanded version of Object Assembly.

An expanded set of puzzles was created for tryout versions of Object Assembly. Included were some more complex objects that matched a set of requirements suggested by neuropsychological research (Dean C. Delis, personal communication, July, 1987). Two classes of objects were developed, those with "interior detail" (printed lines or colored areas) that provided clues to assembly and those with "global form" without interior detail. The goal was to broaden the range of clinical interpretation that could be attached to variations in item performance, but also to increase the range of complexity of items.

Addition of Modernized Artwork

For several subtests (Object Assembly, Picture Completion, Picture Arrangement), completely new, full-color artwork was added. A review of literature on the effects of color in psychological testing indicated that color would heighten attentiveness and the perception of realism in artwork. A controlled experimental study of colored versions of Picture Arrangement indicated that there may be a slight decrease in difficulty and reaction time. But, since all items were being restandardized, this only resulted in the need for a few more difficult items. Feedback obtained during pilot testing and tryouts indicated positive reactions from children and examiners.

Improvements to the Coding Subtest

To ensure that the Coding subtest retained a strong loading on Factor 3, and in response to requests from users, the format and length of the subtest was improved. For left-handed children, more distance between the "key" and the response positions was provided by placing the key near the top of a full sheet of paper (so that the left hand would not block the target key). Both Coding A and B were lengthened to provide more range of possible scoring. The size and clarity of printing was also enhanced for both parts.

Extension of the Normative Range of Subtests

One of the concerns expressed by reviewers of WISC-R was that the norms tend to "bottom out" at age 6 for certain subtests (e.g., Similarities, where a raw score of 2 results in a scaled score of 7) and "top out" at age 16 (e.g., Arithmetic). New items were drafted to extend these scales both downward and upward. Similarities items from WPPSI-R were reviewed and used as a guide for developing WISC-III items. Several additional items tapping counting and number concepts (using pictures) were added at the bottom of the Arithmetic scale, and more difficult multi-step word-problems were added at the top of the scale. Both simpler and more difficult Picture Arrangement sequences were drafted since all the artwork was being redrawn. Additional Block Designs were added. A complex matrix was added to Matrices. Numerous other examples could be listed.

Following pilot testing and field tryouts, item analyses were conducted using conventional and Rasch difficulty indexes. Because items from WISC-R were present, unchanged in several cases, the relative difficulty of the new items could be assessed and a greater range of difficulty achieved. The nature of Rasch difficulties (Wright & Stone, 1979) allows the test developer to more accurately assess the relative scale-values of items at the

extremes of difficulty (i.e., they can be placed at equal intervals on the Rasch scale eventhough their percentage-correct values are nearly indistinguishable).

Minimization of Item Bias

An extensive body of research existed on the issue of item bias on WISC-R (e.g., Cotter & Berk, 1981; Sandoval, Zimmerman & Woo-Sam, 1983). The general consensus of this research was that there was minimal evidence of ethnic bias when the ability level of ethnic groups was taken into account. However, a number of popular/political criticism had been made of specific items, based on face-validity evidence. Therefore, an extensive empirical analysis was conducted using the item-response data from the original normative sample of WISC-R. Three empirical item-bias statistics were computed for all items: Angoff's (1982) delta-plot method, the partial correlation between coded ethnic status and item score holding total IQ constant (Reynolds, Willson, & Jensen, 1984), and the comparison of Rasch difficulty indexes derived from each ethnic subsample. Very few items were identified by two or more of these procedures as having small differential item effects. Thus, a small number of items in Information, Vocabulary and Comprehension were flagged for replacement by new items in WISC-III. Item bias analyses were later repeated in the field tryout (to be repeated in the standardization data analysis).

In addition to empirical item analyses, an item review panel was composed of psychologists with expertise in Black, Hispanic, Asian, and Native American bias studies with WISC-R. This group provided written reports on all items in WISC-R plus draft items for WISC-III. Additionally, staff members of The Psychological Corporation who are Black and Hispanic professionals provided detailed critiques of draft artwork and items.

Supplementary Materials and Tests

One of the requests most often received in marketing surveys and reviews of WISC-III was the call for a co-normed achievement measure. For differential diagnosis of academic difficulties or learning disabilities, an integrated ability-and-achievement battery is clearly an advantage. Thus, a comprehensive, individually-administered achievement battery is currently under development, tentatively named the Wechsler Individual Achievement Test (WIAT).

In addition, an innovative computer-based test interpretive (CBTI) program is under development. The program will provide sophisticated word-processing capabilities as well as extensive hypothesis-generation, similar to the features offered in the newly announced WPPSI-R Writer software (Wechsler, 1990). Instead of a simple, preprogrammed report, the new software will generate numerous interpretive options and statistical indexes,

all of which can be integrated into reports tailored to the individual characteristics of each child and the needs of a particular professional evaluation and setting.

Summary

In summary, numerous empirical and research-based steps were executed in the planning and development of the standardization edition of WISC-III. Although there were numerous improvements, and an obvious need for a substantial number of new items, the major features of WISC and WISC-R were kept intact in WISC-III, for purposes of historical continuity. Also, because the recent revision of the preschool scales, WPPSI-R, included improvements that align with WISC-R (e.g., addition of WPPSI-R Object Assembly), greater downward age-compatibility has been created.

REFERENCES

Angoff, W.H. (1982). Item bias studies based on item difficulty. In R.A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Campione, J.C., & Brown, A.L. (1978). Toward a theory of intelligence: Contributions from research with retarded children. Intelligence, 2, 279-304.

Carroll, J.B. (1987). New perspectives in the analysis of abilities. In R.R.Ronning, J.A. Glover, J.C. Conoley, & J.C. Witt (Eds.) The influence of cognitive psychology on testing. Hillsdale, NJ: Erlbaum.

Carroll, J.B. (1989). Factor analysis since Spearman: Where do we stand? In R. Kanfer, P.L. Ackerman, & R. Cudek (Eds.) Abilities, motivation and methodology. Hillsdale, NJ: Erlbaum.

Cattell, R.B., & Horn, J.L. (1981). Scientific basis of ability testing. American Psychologist, 36, 1012-1020.

Cotter, D.E., & Berk, R.A. (1981). Item bias in the WISC-R using black, white, and hispanic learning disabled children. Paper presented at the meeting of the American Educational Research Association, Los Angeles, April.

Das, J.P., Kirby, J., & Jarman, R.F. (1975). Simultaneous and successive sytheses: An alternative model for cognitive abilities. Psychological Bulletin, 82, 87-103.

Elliott, C.D. (1990). Differential ability scales. San Antonio, TX: The Psychological Corporation.

Embretson, S.E. (Ed.) (1985). Test design: Developments in psychology and psychometrics. New York: Academic Press.

Flynn, J.R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. Psychological Bulletin, 101, 171-191.

Gardner, H. (1982). Frames of mind. New York: Basic Books.

Gustafsson, J.E. (1984). A unifying model for the structure of intellectual abilities. Intelligence, 8, 179-203.

Hunt, E. (1987). Science, technology and intelligence. In R.R. Ronning, J.A. Glover, J.C. Conoley, & J.C. Witt (Eds.) The influence of cognitive psychology on testing. Hillsdale, NJ: Erlbaum.

Jensen, A. (1980). Bias in mental testing. New York: Free Press.

Jensen, A.R. (1987). The g beyond factor analysis. In R.R. Ronning, J.A. Glover, J.C. Conoley, & J.C. Witt (Eds.) The influence of cognitive psychology on testing. Hillsdale, NJ: Erlbaum.

Kaufman, A.S. (1979). Intelligent testing with the WISC-R. New York: Wiley.

Pellegrino, J.W., & Glaser, R. (1979). Cognitive components and correlates in the analysis of individual differences. Intelligence, 3, 187-214.

Reynolds, C.R., Willson, V., & Jensen, A.R. (1984). Black-white differences in simultaneous and sequential processing independent of overall ability differences. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April.

Sandoval, J., Zimmerman, I.R., & Woo-Sam, J.M. (1983). Cultural differences on WISC-R verbal items. Journal of School Psychology, 21, 49-55.

Sattler, J. (1988). Assessment of children. (3rd Ed.) San Diego: Jerome Sattler, Publisher.

Spearman, C. (1927). The abilities of man. New York: Macmillan.

Spring, C., & Farmer, R. (1975). Perceptual span in poor readers. Journal of Reading Behavior, 7, 297-305.

Sternberg, R.J. (1984). What should intelligence tests test? Implications of a triarchic theory of intelligence. Educational Researcher, 13(1), 5-15.

Sternberg, R.J. (in press). Multidimensional ability test. San Antonio, TX: The Psychological Corporation.

Sternberg, S. (1966). High-speed scanning in human memory. Science, 153, 652-654.

Thurstone, L.L. (1938). Primary mental abilities. Psychometric monographs, No. 4.

Vernon, P.E. (1950). The structure of human abilities. New York: Wiley.

Wechsler, D. (1974). WISC-R manual. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1990). WPPSI-R Writer. [Computer program]. San Antonio, TX: The Psychological Corporation.

Wright, B.D., & Stone, M.H. (1979). Best test design. Chicago: MESA Press.

Figure 1

A Wechsler Hierarchical Model
of Intelligence